BMC Bioinformatics

BioMed Central

Methodology article

Open Access

Genetic interaction motif finding by expectation maximization – a novel statistical model for inferring gene modules from synthetic lethality

Yan Qi[1,2], Ping Ye[1,2] and Joel S Bader*[1,2]

Address: [1]Biomedical Engineering Department, Johns Hopkins University, North Charles Street, Baltimore, MD, 21218, USA and [2]High-Throughput Biology Center, Johns Hopkins School of Medicine, 733 North Broadway, Baltimore, MD 21205, USA

Email: Yan Qi - yanqi@jhu.edu; Ping Ye - pingye@bme.jhu.edu; Joel S Bader* - joel.bader@jhu.edu

* Corresponding author

Published: 06 December 2005

BMC Bioinformatics 2005, **6**:288 doi:10.1186/1471-2105-6-288

Received: 26 July 2005

Accepted: 06 December 2005

This article is available from: http://www.biomedcentral.com/1471-2105/6/288

Abstract

Background: Synthetic lethality experiments identify pairs of genes with complementary function. More direct functional associations (for example greater probability of membership in a single protein complex) may be inferred between genes that share synthetic lethal interaction partners than genes that are directly synthetic lethal. Probabilistic algorithms that identify gene modules based on motif discovery are highly appropriate for the analysis of synthetic lethal genetic interaction data and have great potential in integrative analysis of heterogeneous datasets.

Results: We have developed Genetic Interaction Motif Finding (GIMF), an algorithm for unsupervised motif discovery from synthetic lethal interaction data. Interaction motifs are characterized by position weight matrices and optimized through expectation maximization. Given a seed gene, GIMF performs a nonlinear transform on the input genetic interaction data and automatically assigns genes to the motif or non-motif category. We demonstrate the capacity to extract known and novel pathways for Saccharomyces cerevisiae (budding yeast). Annotations suggested for several uncharacterized genes are supported by recent experimental evidence. GIMF is efficient in computation, requires no training and automatically down-weights promiscuous genes with high degrees.

Conclusion: GIMF effectively identifies pathways from synthetic lethality data with several unique features. It is mostly suitable for building gene modules around seed genes. Optimal choice of one single model parameter allows construction of gene networks with different levels of confidence. The impact of hub genes the generic probabilistic framework of GIMF may be used to group other types of biological entities such as proteins based on stochastic motifs. Analysis of the strongest motifs discovered by the algorithm indicates that synthetic lethal interactions are depleted between genes within a motif, suggesting that synthetic lethality occurs between-pathway rather than within-pathway.

Background

Much recent research efforts have been devoted to studying gene functions in the context of highly dynamic and modular cellular networks [1-4]. Valuable information about a gene's function can be obtained from its interaction with other genes [5]. Apart from the traditional hierarchical way of gene function annotation, functional genomics takes a bottom-up approach to assemble gene interaction networks based on all pair-wise gene interactions detected. From such genetic interaction maps, Functional modules representing various biological pathways and processes can then be extracted by computational approaches. Those modules naturally suggest novel gene functions in the relevant biological processes [6]. The interactions between genes are of course highly dynamic spatially and temporally. However, one of the most intuitive yet fundamental questions about genetic interactions is whether the normal functioning of two genes depends on each other. Synthetic lethality identifies genes that complement each other's function: two genes are synthetic lethal if either single mutant is viable, but the double mutant combination is lethal. High-throughput experiments such as synthetic genetic array (SGA) [7] and synthetic lethality analyzed by microarray (SLAM) [8,9] have been done for genome-wide synthetic lethality analysis on *Saccharomyces cerevisiae*, where a single mutant (query gene) is introduced into the complete pool of viable yeast single-deletion (library gene) strains. Synthetic lethality data obtained through SGA, SLAM or RNA interference has shed much new light on essential biological pathways and the function assignment for many previously uncharacterized genes for the model organisms yeast and C. elegans [10,11]. Hierarchical clustering of the SGA dataset suggest that two synthetic lethal genes are likely to (i) reside in two redundant parallel pathways or (ii) complement each other's function in two branches of one essential pathway [12]. Computational methods integrating physical protein interactions and other genomic features seem to suggest that significantly more synthetic lethal interactions occur between parallel pathways [13,14]. Given the incomplete and error-prone synthetic lethal interaction map, it is highly desirable to investigate methods that extract biologically relevant information probabilistically, which accommodates network properties such as degree distribution and confidence of the links. Along this line, we have developed in this study a probabilistic model for characterizing synthetic lethal interaction motifs and an algorithm that automatically groups genes sharing similar motifs into pathways. When applied to the SGA dataset, our method automatically uncovers known and novel gene modules that correlate favourably with Gene Ontology (GO) annotations.

Results

Data sources

Genetic interaction data is obtained from SGA analysis in yeast [12]. The original query gene set includes 126 non-essential genes and 6 essential genes, tested against a library of all non-essential gene deletions. Interpretation of synthetic lethality involving essential genes is problematic since the intermediate (viable) phenotypes exhibited by conditional alleles of essential genes may include loss of function, unregulated function, and gain of function aspects. Thus we focus on synthetic lethal interactions between null alleles of non-essential genes, which by definition result from solely loss of function mutations. Ignoring library genes that have no interaction with any of the 126 query genes, our dataset consists of 126 query genes linked to 982 library genes by 4287 interactions. Both the query and the library sets contain hubs with high interaction counts (Supp. Figs. S3, S4, and S5).

Yeast protein complex data were obtained from two high-throughput studies, TAP and HMS-PCI [15,16]. Protein complexes that contained two or more non-essential proteins were used (353 complexes from TAP and 427 complexes from HMS-PCI).

Computational method

The Expectation maximization (EM) algorithm has been widely used to detect motifs in biopolymer sequences, where a position weight matrix representing a recurring pattern (such as DNA binding sites or promoter regions) in multiple unaligned sequences is built iteratively by maximum likelihood scoring [17-20]. Such probabilistic approach is most suitable for the detection of patterns with a stochastic nature, which we have little prior knowledge of. In this study, we have developed an algorithm for finding genes in the same pathway, which we shall refer to as Genetic Interaction Motif Finding by expectation maximization (GIMF). Note the difference between motif here defined by genetic interaction pattern and the network topological motifs [21]. The model is developed under the hypothesis that genes within the same pathway exhibit a similar pattern of synthetic lethal interactions where a subset of common interaction partners are genes in complementary pathways [12-14]. For example, RVS161 and RVS167 are two queries that belong to the RVS161 complex. Enhanced synthetic lethal interactions with members of the RPD3 complex have been observed (Fig. 1). The RVS161 complex proteins are AR adaptor proteins involved in actin regulation, endocytosis and viability following starvation or osmotic stress. The RPD3 histone deacetylase complex is involved in silencing at telomeres. In particular, DEP1, a member of the RPD3 complex is a transcriptional modulator of phospholipids biosynthesis and also maintains mating efficiency and sporulation. Thus it is reasonable to infer that these two

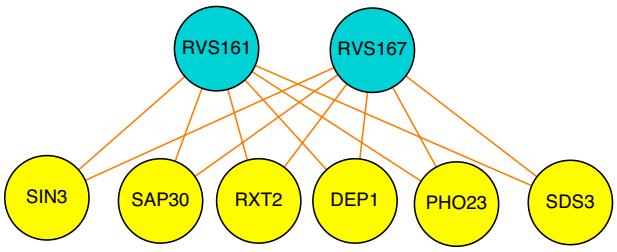


Figure 1
Synthetic lethal interactions between complementary pathways. Proteins in the RVS161 complex (RVS161, RVS167) have enriched synthetic lethal interactions (orange lines) with proteins in the RPD3 complex. The RVS161 complex and RPD3 complex are associated with endocytosis/viability following starvation and telomere silencing, respectively.

protein complexes are functionally complementary during endocytosis and mating or sporulation after starvation when the biological processes of the two complexes are tightly coupled.

In our analysis, we focus on finding motifs from the synthetic lethal interaction patterns of query genes. Let $X_i = [X_{i1} \ldots X_{iN}]$ denote the interaction partner list for query gene i, where $X_{ij} = 1$ if i interacts with library gene j and $X_{ij} = 0$ otherwise. Thus the entire data set is X_i, $i = 1,2,\ldots,Q$. The total numbers of query is Q = 126 and the total number of the library genes that interact with at least one query gene is $N = 982$. We initiate a search with a query gene s and aim to find all other genes in the same pathway as the seed gene s. We do this by iteratively constructing a motif for the group and hence identifying motif members.

Mathematically, we divide the query gene set into two sets, a motif set $A = \{A_i\}$, $i = 1,2,\ldots a_M$, initialized to contain just the seed gene, and a non-motif set $B = \{B_i\}$, $i = a_M+1$, a_M+2, L, a_M+b_M, containing the remaining genes. The number of query genes in the motif and non-motif sets are a_M and b_M, respectively, with $a_M+b_M=Q$. We assume that genes in the motif set and those in the non-motif set have different probabilities of interacting with a library gene j, which are denoted by p_{aj} and p_{bj}, respectively. As will be explained in DISCUSSION, this allows existence of hub library genes explicitly. The probability that query i belongs to the motif set is denoted by z_i. The parameters p_{aj}, p_{bj} and z_i, where $j = 1,2\ldots,N$ and $j = 1,2\ldots,Q$, are estimated iteratively.

The expectation maximization (EM) algorithm has been used for maximum likelihood estimation with missing information. In our scenario, given a seed gene, missing information is represented by the correct partition of the entire gene pool into a motif set A and a non-motif set B starting from an initial motif estimate provided by the seed. The likelihood function, i.e. the conditional probability of observing measured data given the partition, is

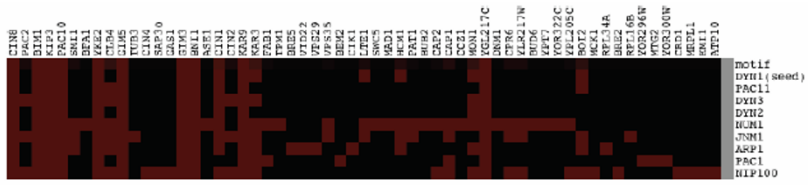
$$L = P(X \mid A, B) = \prod_{i=1}^{a_M}\prod_{j=1}^{N}\left[X_{ij}p_{aj} + (1-X_{ij})(1-p_{aj})\right] \times \prod_{i=a_M+1}^{a_M+b_M}\prod_{j=1}^{N}\left[X_{ij}p_{bj} + (1-X_{ij})(1-p_{bj})\right]. \quad (1)$$

Thus the log likelihood function is

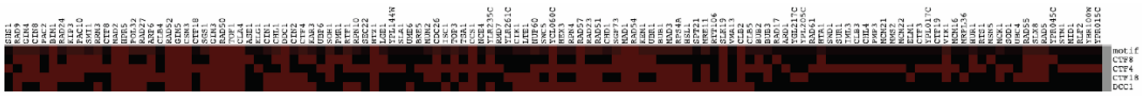
$$\begin{aligned}\log L &= \sum_{j=1}^{N}\sum_{i=1}^{a_M}\left[X_{ij}\log p_{aj} + (1-X_{ij})\log(1-p_{aj})\right] \\ &+ \sum_{j=1}^{N}\sum_{i=a_M+1}^{a_M+b_M}\left[X_{ij}\log p_{bj} + (1-X_{ij})\log(1-p_{bj})\right] \\ &= a_M\sum_{j=1}^{N}\left[f_{aj}\log p_{aj} + (1-f_{aj})\log(1-p_{aj})\right] \\ &+ b_M\sum_{j=1}^{N}\left[f_{bj}\log p_{bj} + (1-f_{bj})\log(1-p_{bj})\right]\end{aligned} \quad (2)$$

where $f_{aj} = \frac{1}{a_M}\sum_{i=1}^{a_M} X_{ij} = \frac{n_{aj}}{a_M}$ and $f_{bj} = \frac{1}{b_M}\sum_{i=a_M+1}^{a_M+b_M} X_{ij} = \frac{n_{bj}}{b_M}$ are the observed frequencies of the interaction between the library gene j and query genes in motif set and non-motif set, respectively. Maximizing the log likelihood function, we obtain the maximum likelihood estimates for p_{aj} and p_{bj}, which are simply the sample frequencies, i.e. $\hat{p}_{aj} = f_{aj}$ and $\hat{p}_{bj} = f_{bj}$ (unless otherwise stated, an overlying hat denotes the maximum likelihood estimate of a parameter). These estimates cannot be obtained when the partition information is missing. The EM algorithm starts with an initial guess of the solution and iteratively updates the model parameters with expected information obtained by maximum likelihood estimation in each step. More specifically, each iteration comprises two steps, an expectation (E) step and a maximization (M) step.

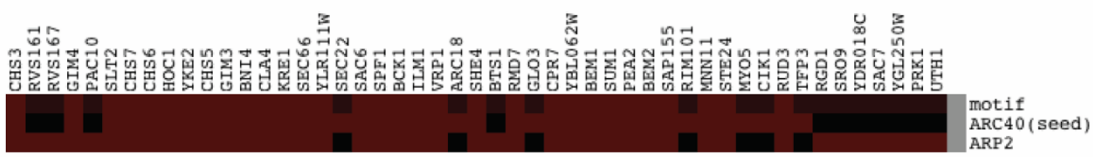
Let us assume that q iterations have been completed. At the start of the E-step of iteration q+1, the estimates for the model parameters from the M step of the previous iteration, $\hat{p}_a^{(q)}$, $\hat{p}_b^{(q)}$ and $\hat{z}_i^{(q)}$, are available. Let Y_i be a motif indicator, i.e. $Y_i = 1$ if gene i belongs to the motif set and $Y_i = 0$ otherwise. Then the conditional probability of



A. Dynein Dynactin Pathway



B. Sister Chromatid Cohesion



C. Actin Patch Assembly

Figure 2

Representative genetic interaction patterns of the seed, the motif and motif members. Seed genes for the motifs are (a) DYN1 (b) CTF8 (c) ARC40. The columns correspond to library genes (interaction partners of query genes). Library genes that have no interaction with the seed and the motif members are not shown. Synthetic lethal interactions are represented by red squares. The non-binary values of the motifs are shown by intermediate colors changing from black to red.

observing $\bar{X}_i$, given that gene *i* belongs to the motif set and $\hat{p}_a^{(q)}$, is

$$\mathrm{P}(X_i \mid Y_i = 1, p_a^{(q)}) = \prod_{j=1}^{N} \left[p_{aj}^{(q)} X_{ij} + (1 - p_{aj}^{(q)})(1 - X_{ij}) \right]. \quad (3)$$

Similarly the conditional probability of observing $\bar{X}_i$ given that gene *i* belongs to the non-motif set and $\hat{p}_b^{(q)}$, is

$$\mathrm{P}(X_i \mid Y_i = 0, p_b^{(q)}) = \prod_{j=1}^{N} \left[p_{bj}^{(q)} X_{ij} + (1 - p_{bj}^{(q)})(1 - X_{ij}) \right]. \quad (4)$$

By Bayes formula, the probability that a gene *i* belongs to the motif set given observed data and current model estimates is,

$$\begin{aligned} \hat{z}_i^{(q+1)} &= P(Y_i = 1 \mid X_i, \hat{p}_a^{(q)}, \hat{p}_b^{(q)}) \\ &= \frac{P(X_i \mid Y_i = 1, \hat{p}_a^{(q)}) P_0(Y_i = 1)}{P(X_i \mid Y_i = 1, \hat{p}_a^{(q)}) P_0(Y_i = 1) + P(X_i \mid Y_i = 0, \hat{p}_b^{(q)}) P_0(Y_i = 0)}, \end{aligned} \quad (5)$$

where $P_0(Y_i = 1) = a_M^{(0)} / Q$ is the prior probability that gene *i* belongs to the motif set.

The expected number of interactions with a library gene *j* is the weighted sum of all the query genes' interactions with gene *j*, where X_{ij} is weighted by $z_i^{(q+1)}$, *i = 1, 2, L, N*. These expected numbers ε_{aj} and ε_{bj} for motif and non-motif query genes for iteration *q* + 1 are

$$\begin{aligned} \varepsilon_{aj}^{(q+1)} &= \mathrm{E}\left(n_{aj} \mid X, p_{aj}^{(q)} \right) = \sum_{i=1}^{Q} z_i^{(q+1)} \cdot X_{ij} \ ; \\ \varepsilon_{bj}^{(q+1)} &= \mathrm{E}\left(n_{bj} \mid X, p_{bj}^{(q)} \right) = \sum_{i=1}^{Q} \left(1 - z_i^{(q+1)} \right) \cdot X_{ij} \ . \end{aligned} \quad (6)$$

In the M step, model parameters are updated with expected numbers,

$$\begin{aligned} \hat{p}_{aj}^{(q+1)} &= \varepsilon_{aj}^{(q+1)} / \hat{a}_M^{(q+1)}; \\ \hat{p}_{bj}^{(q+1)} &= \varepsilon_{bj}^{(q+1)} / \hat{b}_M^{(q+1)}; \end{aligned} \quad (7)$$

where $\hat{a}_M^{(q+1)} = \sum_{i=1}^{Q} z_i^{(q+1)}, \hat{b}_M^{(q+1)} = Q - \hat{a}_M^{(q+1)}$.

Convergence of the algorithm is assessed by $|t^{(q+1)} - t^{(q)}| < 10^{-4}$ for all of the model parameter estimates $\hat{p}_a$, $\hat{p}_b$ and $\hat{z}$.

Given a seed gene *s*, the model parameters are initialized as follows:

$$\hat{p}_{bj}^{(0)} = Q^{-1} \sum_{i=1}^{Q} X_{ij} \qquad (8)$$

$$\hat{p}_{aj}^{(0)} = p \cdot X_{sj} + \hat{p}_{bj}^{(0)} \cdot (1 - X_{sj}). \qquad (9)$$

While the sum over query genes to estimate the initial background probability $\hat{p}_{bj}^{(0)}$ should formally exclude the seed gene, we include it to prevent initialization to zero probability in the case that only the seed gene has an interaction with library gene *j*. The choice of $0 < p < 1$ in Eq. (9) depends on our confidence about the seed gene's interactions. We have used $p = 0.95$ for the analysis that follows, corresponding to a false positive rate of 5% in the SGA experiment. The motifs extracted are not sensitive to the choice of *p* in the vicinity of the experimental false positive rate (e.g. $p \geq 0.9$). A more detailed discussion on motifs' dependencies on *p* will be given in the DISCUSSION section and it will be made clear that by adjusting *p*, it is possible to systematically retrieve motif members based on the degree of similarity between their genetic interaction patterns and that of the seed. The number of motif members for any seed gene is most likely to be a small portion of the size of the query gene set Q = 126. Thus the number of genes in the motif set is initialized by $\hat{a}_M^{(0)} \in [5,15]$. Our analysis shows that the algorithm is not sensitive to the choice of $\hat{a}_M^{(0)}$ in this range.

Results on SGA dataset

Outputs of our model are $\hat{p}_a$, $\hat{p}_b$ and $\hat{z}_i$. The motif and background interaction probabilities $\hat{p}_a$ and $\hat{p}_b$ are two position weight matrices with continuous elements in the range of [0,1]. Unlike in the case of DNA binding site detection, the converged probabilities $\hat{z}_i$ computed for the SGA dataset are either very close to 0 or very close to 1; intermediate values have not been observed. Thus, given a seed gene, the remaining genes are naturally categorized as motif genes (with $z_i \approx 1$) or non-motif genes (with $z_i \approx 0$). We call motif genes motif members of the seed. Three representative motifs are shown along with the genetic interaction patterns of the seed genes *DYN1*, *CTF8* and *ARC40* and their motif members (Fig. 2). Table 1 shows the groups of motif members obtained for seed genes *ARL1*, *SKT5*, *CTF8* and *RIC1* when various values of *p* are used. Motif members obtained with $p = 0.95$ for 13 seed genes are listed in Table 2. The full table is available as supporting material (additional file gimf-motifs.txt). The seven groups of genes thus identified agree with groups observed with hierarchical clustering [12], which supports GIMF's capacity in extracting biologically relevant gene pathways.

One important property of GIMF is that it is non-commutative: if gene A identifies gene B as a motif member, it is not necessarily true that gene B identifies gene A as its motif member. Interestingly, we have observed that a seed gene tends to first pull up motif members that share a globally similar interaction pattern. If such genes are lacking, then it finds genes with locally similar interaction pattern. This enables us to probe the case when two genes' interaction partners are only similar on a local scale. This is not

Table 1: Motif members of four seed genes.

Seed	Motif members			
	p = 0.6	*p* = 0.7	*p* = 0.8	*p* = 0.95
ARL1	ARL3, SWF1, RIC1, YPT6		RIC1, YPT6, ARL3	
SKT5	CHS6, CHS3, CHS7, CHS5	CHS3, CHS7, CHS5		CHS3, CHS5
CTF8	CTF4, CTF18, DCC1, BIM1, CIN8, KAR3	CTF4, CTF18 DCC1, BIM1	CTF4, CTF18, DCC1	CTF4, CTF18, DCC1
RIC1	YPT6			

Motif members of ARL1, SKT5, CTF8 and RIC1 are obtained by choosing different values of initialization parameter *p*. Most motifs show little dependency on p for $p \geq 0.9$.

Table 2: Seven representative motifs identified by GIMF.

	Pathway or complex	Seed gene	Motif gene list
1	Actin patch assembly	ARC40	ARP2
2	Chitin synthase III pathway	CHS7	CHS3, SKT5, CHS5
		CHS6	CHS3, SKT5
3	Prefoldin complex	PAC10	GIM3, GIM4, GIM5, YKE2
4	Membrane traffic	ARL1	ARL3, RIC1, YPT6
		GYP1	RIC1
5	Dynein Dynactin pathway	DYN1	ARP1, DYN1, PAC11, YMR299C, DYN2, JNM1, PAC1, NIP100, NUM1
		PAC1	ARP1, DYN1, PAC11, YMR299C, DYN2, JNM1, NIP100, NUM1
		JNM1	ARP1, DYN1, PAC11, YMR299C, DYN2, NIP100, NUM1
		NUM1	JNM1
6	DNA replication checkpoint	MRC1	TOF1
7	Sister chromatid cohesion	DCC1	CTF4, CTF18
		CTF8	CTF18, DCC1, CTF4

These gene modules correlate well with pathways inferred by hierarchical clustering (Tong et al. 2004).

possible with pair-wise comparison metrics, which are commutative. For a more systematic analysis, we use GIMF to build gene networks. First, query genes with very few interactions (5 or fewer) are removed from the list of seeds. Then each of the remaining query genes is used as a seed and its motif members are generated by GIMF. For every query gene pair(i, j), if i and j are each other's motif member, then connect i and j with a Type 1 edge. We call the network thus constructed a Type 1 GIMF network (Fig. 3). This network contains 31 nodes and 42 edges, which form two clusters and eight individual pairs. The smaller cluster is a fully connected sub-graph corresponding to the PAC10 complex. The larger cluster with 10 genes (ARP1, NUM1, DYN1, PAC11, PAC1, DYN2, JNM1, YMR299C, NIP100, KIP2), representing the Dynein-Dynactin spindle orientation pathway. KIP2 was not detected by hierarchical clustering [12].

Apparently, the bi-directional rule only retains genes with globally similar interaction pattern. This can be quite stringent since genes have multiple functions and two genes operating in one pathway may have distinct roles in other pathways they participate and thus only share a fraction of synthetic lethal interaction partners. Thus we extend Type 1 network by the following simple rule: for each gene pair (*i*, *j*) in the Type 1 network, add common motif members *k* of genes *i* and *j* that are not already in the network (hence neither *i* nor *j* is motif member of *k*). Connect *k* to *i* and *j* with a Type 2 edge. We call the extended network a Type 2 network (Fig. 4). This analysis reveals more information in the *Dynein-Dynactin* pathway. The majority of Type 2 edges occur between the group members of this cluster, which elevates the confidence that the genes within this cluster are closely related. Evidence that genes in this cluster are biologically related include the presence of a dynactin protein complex (*ARP1*, *JNM1*, *NIP100*), reported protein-protein interactions between *NIP100-PAC11*, *PAC11-DYN2*, *PAC11-NUM1* [22,23] and the suggestion that *YMR299C* functions as dynein light intermediate chain [12]. In the Type 2 network, several new members are incorporated into the cluster, including *NBP2*, *BIK1* and *CTF18*. The molecular function of *NBP2* and *CTF18* are unknown while *BIK1* is involved in microtubule binding. *NBP2* shows hyperosmotic and heat response and is a negative regulator of protein kinase activity. *CTF18* is a subunit of a complex with *CTF8P* that shares some subunits with Replication Factor C and is required for sister chromatid cohesion. It has been known that the mutants of six genes (*NUM1*, *DYN1*, *DYN2*, *ARP1*, *JNM1*, *NIP100*) in this cluster show nuclear migration defect in cell division process. A recent experiment has confirmed that deletion mutants of *KIP2*, *BIK1* and *CTF18* also exhibit moderate to severe nuclear migration defects [24]. These three genes have not been detected by two way clustering [12].

Under our hypothesis, genes with a similar synthetic interaction pattern (especially when the similarity is global) are likely to reside in the same pathway or map to proteins in the same complex. Thus the motif members are expected to have functional similarities at various levels. We evaluate the biological relevance of the Type 1 and Type 2 networks by computing three parameters for each edge (gene pair): the correlations with the Gene Ontology (GO) annotations (described in Appendix); the fraction of gene products that are within the same protein complex as determined by high-throughput mass spectrometry; and the fraction that are synthetic lethal. These parameters have also been computed for all directly synthetic lethal gene pairs. The Type 1 gene pairs' correlations for biological process, molecular function and cellular component GO annotations are (0.47, 0.20, 0.43), while those of the Type 2 network are (0.47, 0.15, 0.40), comparing to (0.25, 0.05, 0.31) for directly synthetic lethal gene pairs (Table 3). Clearly, much tighter functional associations are obtained between gene pairs with either globally or

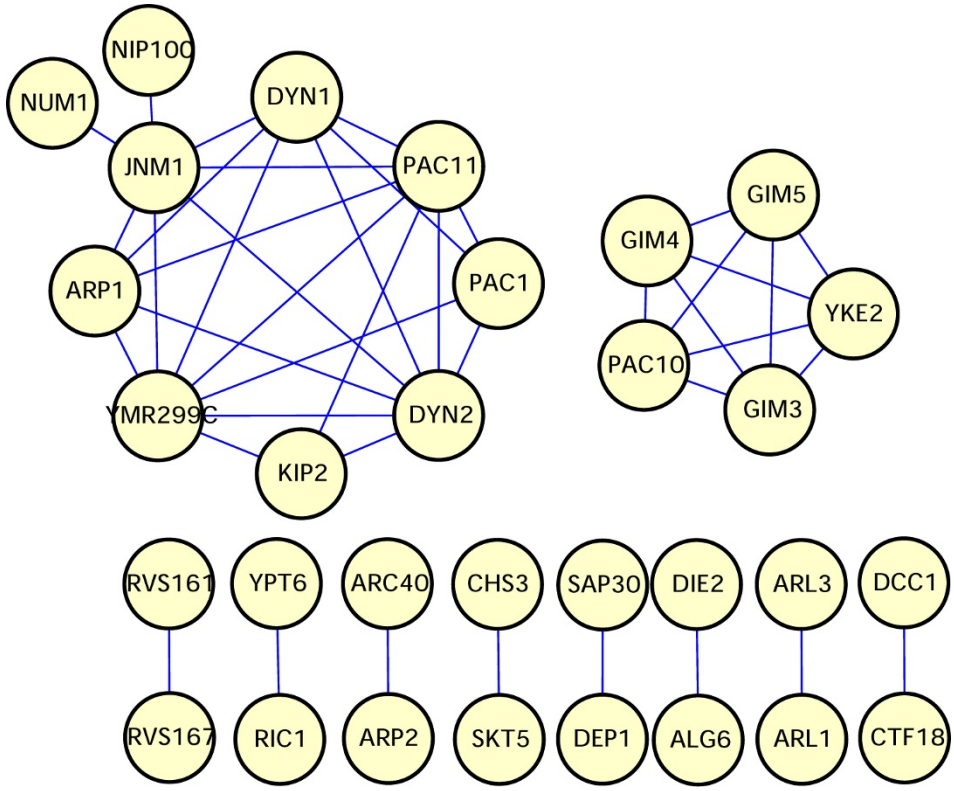


Figure 3
GIMF Type I network. The network is created by applying Rule 1 to the motif member lists of all query genes. The network contains 31 nodes and 42 edges, where the nodes are query genes and an edge between node *i* and node *j* indicates that *i* and *j* are each other's motif members.

locally similar synthetic lethal interactions than gene pairs that are directly synthetic lethal interactions, confirming the observation of between-pathway enrichment by Wong et al. and Kelly et al. [13,14]. Significantly more Type 1 gene pairs map to proteins within the same complex than either Type 2 gene pairs or directly synthetic lethal gene pairs. Same-complex membership may explain the higher molecular function correlation for Type 1 gene pairs.

Discussion

In this section, we explore a few important issues in terms of the robustness and tuning of GIMF. Without loss of generality, the discussion is primarily based on learning pathway association on the SGA dataset.

It has been widely known that EM algorithm very often converges to local maxima in the evaluation of posterior likelihood function or log-likelihood [18,19]. In application to motif (e.g. transcription binding sites) discovery in DNA sequences, early versions EM assumed the existence of a single motif and aimed to find the motif that globally optimized the likelihood function. However, when multiple consensus sequences are present in the dataset, numerous local maxima in the likelihood function can well correspond to biologically meaningful motifs. One approach to finding multiple motifs is to initialize the EM from different starting points, typically selected from patterns occurring in the data, which may then relax to local maxima. This approach may be enhanced, as in the MEME algorithm, by erasing motifs previously found so that multiple motifs are found in decreasing order of likelihoods. Using these two strategies, MEME successfully detects multiple promoter consensuses from the combined CRP/LexA datasets[18].

In GIMF, we achieve a similar effect by initializing the model using seed gene's interactions, thus narrowing down the search space to the module that includes the seed. Without any prior knowledge of goodness of seeds and their consensus interactions, two problems are noteworthy: i) Motifs generated by different seeds may be redundant; ii) Certain motifs may deviate from their seeds during the iterative process. These two issues are addressed below:

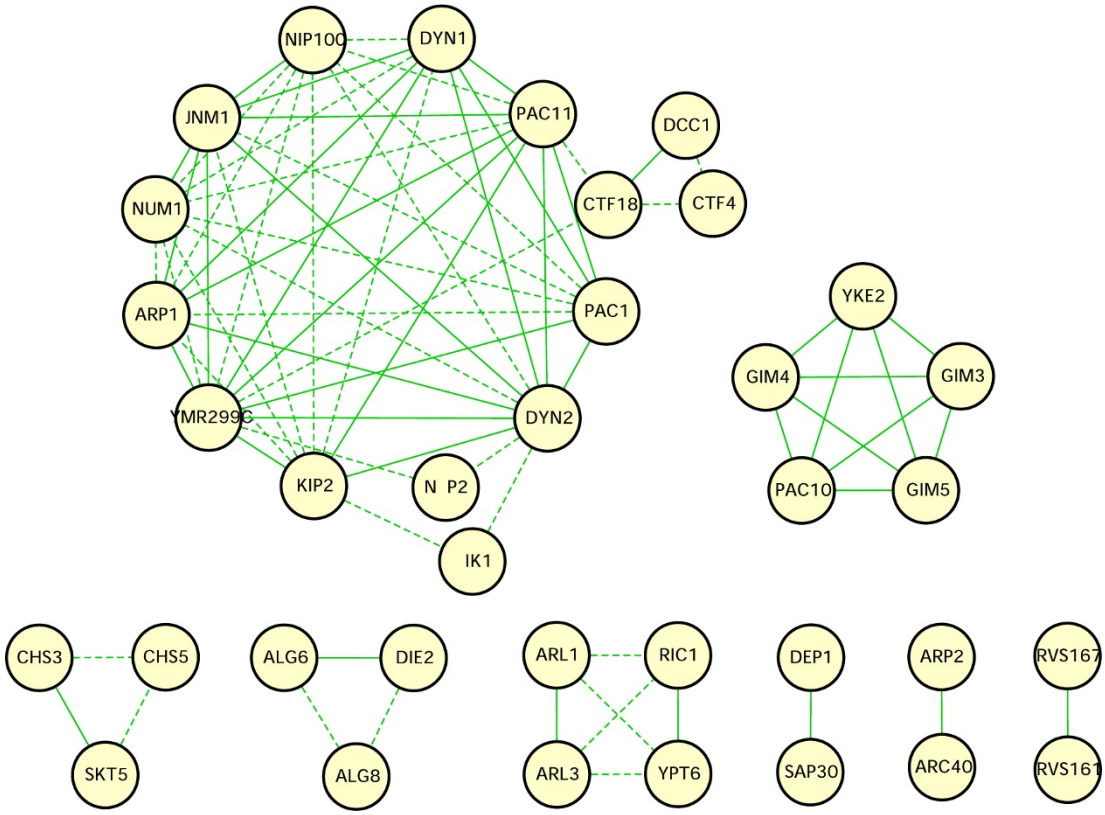


Figure 4
GIMF Type 2 network. This network is created by applying Rule 2 to all the edges in the GIMF Type 1 network. See text for details. The solid edges are inherited from the Type 1 network while the dashed edges are added by applying Rule 2.

i) Motifs generated by different seeds may be redundant

To better understand the dissimilarity between distinct motifs, we have calculated the Euclidean distance between each pair of motifs $\hat{p}_{al}^{S1}$ and $\hat{p}_{al}^{S2}$ generated by seed S1 and S2, respectively, resulting in a 126 by 126 distance matrix D. To visualize this matrix, we embed it in two dimensions using classic multidimensional scaling, which is essentially equivalent to projecting the two leading principal components. Motifs corresponding to connected components in the Type 1 network are close together in this embedding (Fig. 5). In most cases, the seed genes in a Type 1 connected component have either overlapping motifs (several motifs collapse onto one point) or motifs that are very similar to each other. In comparison with the greater number of maxima identified for all query genes (Fig. S2), this analysis suggests that the local maxima corresponding to queries in the Type 1 network are reproducibly identified. These local maxima could be considered global maxima conditioned on the seed gene remaining in the motif.

ii) Certain motifs may deviate from their seeds during the iterative process

In some cases, the EM algorithm may eject a seed gene from a motif. This occurs for eight seed genes when p = 0.95 using the threshold $Z_i > 0.9$ (Table S2). Those seeds either have few interactions and/or have interactions that overlap largely with the interaction partners of some hub genes, such as the *PAC10* complex genes. Indeed, most of their motif members are hub genes, whose interaction profiles override that of the seed genes during the iteration. Thus to ensure each seed stay in the motif, we can slightly modify the algorithm by fixing $Z_{seed} = 1$ during all iterations. In other words, the motif search is conditioned

Table 3: GO annotation correlations for GIMF Type 1, Type 2 gene pairs, and gene pairs that are directly synthetic lethal (SL).

Gene pairs	GO correlation			FSL	FPC	Number of pairs	Number of genes
	P	F	C				
Type 1	0.47	0.20	0.43	0	0.26	42	31
Type 2	0.47	0.15	0.40	0.07	0.14	78	36
SL	0.25	0.05	0.31	--	0.01	3474	1004

P: biological process; F: molecular function C: cellular component. FSL: fraction of pairs that are directly synthetic lethality; FPC: fraction of pairs that are within the same protein complex. The GIMF Type 1 network contains 31 genes and 42 edges while the Type 2 network contains 36 genes and 78 edges.

on the seed being part of the motif. Indeed, for the eight seeds mentioned above, this modification keeps the seed gene itself in the motif till convergence while all other motif members stay unchanged. Clearly, this procedure has no effect on the 106 seeds that are already in motif without such conditioning.

Symmetry imposed by Type 1 edges serves as a conservative filtering procedure that eliminates redundancy and impact of hub genes dominating interaction profiles, which reveals gene networks with tight functional correlations, which supports our finding that the local optimums in GIMF correspond to biologically relevant modules.

We have investigated how the choice of p, the initialization parameter that represents our confidence on seed gene's interactions, affect the motifs. Indeed, the sensitivities of different motifs to p is non-uniform. We quantify the goodness of a seed and its motif by observing stability of its motif members across different choices of p. Genes with less than five interaction partners (12 out of 126) are not used as seeds. For every remaining query gene, we extract its motif members with p ranging from 0.6 to 0.95. The sets of motif members extracted at $p = 0.95$ is used as the reference to compute a Jaccard coefficient [14]. Denote the set of motif members for seed gene i obtained with initialization parameter p by M_p^i. The corresponding Jaccard coefficient J_p^i is given by $J_p^i = \frac{M_p^i \cap M_{p=0.95}^i}{M_p^i \cup M_{p=0.95}^i}$. The seeds can be divided into four categories based on their Jaccard coefficient averaged over the different values of p: i) $\overline{J}^i = 1$. Those seeds have invariant motif members and are hence categorized as very strong seeds. Special cases are seeds that exact only themselves. Those seeds have unique interaction patterns but most likely do not have any pathway members included in the query set. ii)$0.9 < \overline{J}^i < 1$. These are strong seeds with almost invariant motif members. iii)$0.6 < \overline{J}^i < 0.9$. This case corresponds to moderately strong seeds whose motif members change moderately and hierarchically. Decreasing p decreases the confidence in the seed gene's interactions, and genes with more distant interaction profiles can be incorporated into the motif set. The motif members converge at confidence level close to the true experimental false positive rate. iv) $\overline{J}^i < 0.6$. Those seeds have highly variable motif members and hence are weak seeds. The numbers of seeds in the four categories are 14, 6, 45 and 49, respectively (Table S1). When analyzing the SGA dataset, $p = 0.95$ is reasonable because interactions in the SGA dataset have been experimentally validated and has a low false-positive rate. This analysis has three important indications: i) Not all the query genes are good seeds, partly due to the incompleteness of the synthetic lethal genetic interaction map in the query axis; ii) To achieve optimal detection of motifs for different seeds, we might need to employ different initialization parameter p. Given a minimum Jaccard coefficient, the algorithm can be optimally initialized for each seed. iii) The Type 1 network obtained at $p = 0.95$ is most likely conservative. Thus to build gene networks with better confidence, we may eliminate bad seeds, relax confidence constraints on strong and moderately strong seeds while imposing an initialization parameter close to 1 on weak seeds.

To better evaluate the statistical significance of motifs detected by GIMF, we have computed the false positive rates on randomized datasets with the same degree distribution as the original synthetic lethal dataset. Randomization is done by a rewiring procedure as detailed in [21]. The fraction of overlapping links between the randomized network and the original network is around 15%. Since a random network should not contain any biologically relevant motif, any motif detected is a false positive. Thus for the GIMF algorithm, we use every query as a seed gene and any motif member returned other than the seed itself is considered a false positive. The numbers of false positives

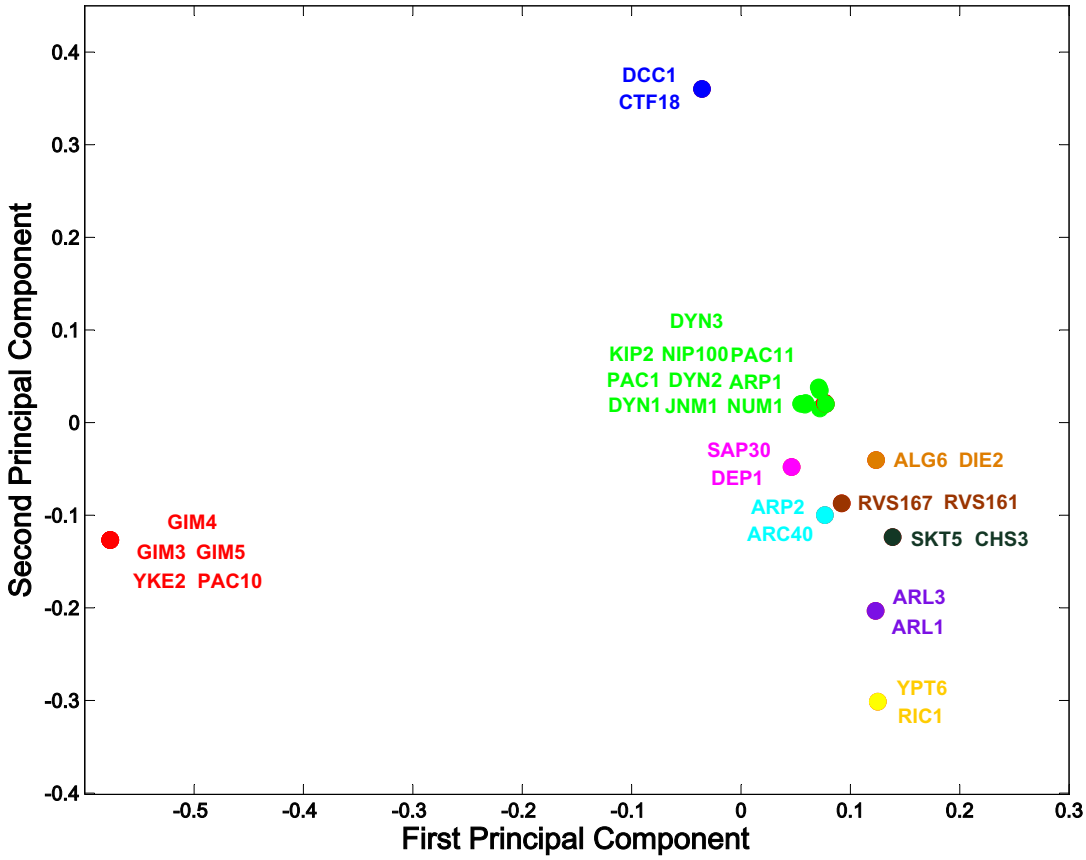


Figure 5
Two-dimensional embedding of Type I network motifs. Each motif is represented by two leading principal components generated by multidimensional scaling based on a Euclidean distance matrix. Colors indicate connected components from the Type I network.

on 100 randomized networks are shown in Fig. S1. Without imposing the bi-directionality constraint, the average total number of false positives for 126 seed genes is 15. The average number of seeds that generates any false positives is 9.7 out of 126. On the real dataset, the number of seeds leading to motif detection is $T = 82$. Thus this corresponds to a p-value of 10^{-15} calculated as tail probability at $T = 82$ from a Poisson distribution. A detailed look at the false positive pairs of GIMF shows that most seeds that lead to false positives have very few interactions with the library genes. The top 10 seeds producing the most false positives have 6.3 interactions on average and their false positive motif genes are mostly promiscuous hub genes. However, no false positives are observed when the promiscuous genes are used as seeds. When bi-directionality is imposed on motif detection, false positive drops to 0 for all the 100 trials. Thus for an asymmetric metric like GIMF, we can impose symmetry constraint to mask the effects of promiscuous genes. Additional information can be obtained by elevating stringency once the reliable gene pairs are identified.

The treatment of hub genes is a problematic issue in the analysis of power-law networks. Hubs arise from many different sources including intrinsic error in the experimental technique (such as sticky proteins in yeast two

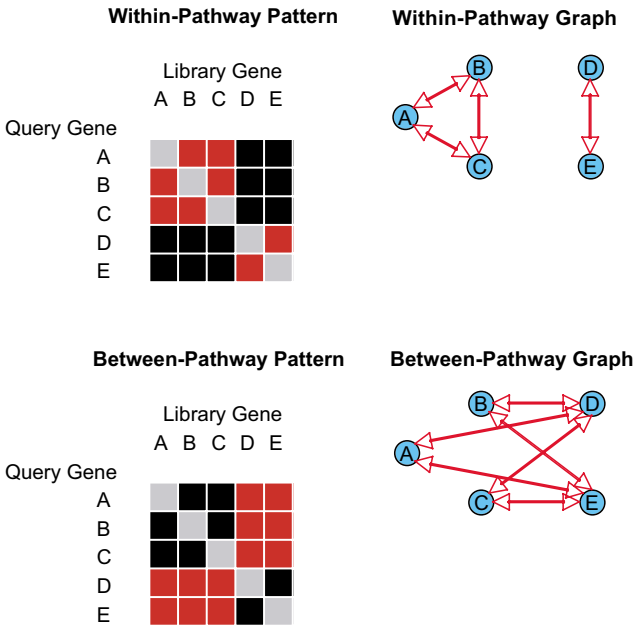


Figure 6
Within-pathway and between-pathway patterns. Partial interaction profiles are shown for five query genes where "interaction", "no-interaction", "prohibited self-interaction" are represented by red, black and grey in the interaction matrix respectively. Genes A, B, C and D, E belong to two different pathways. Interactions involving members of the same pathway are enriched in the within-pathway model and depleted in the between-pathway model.

hybrid system) and experimental bias (such as the choice of query genes for SGA). Because of this heterogeneity, the treatment of promiscuous genes should be context-based. In the case of experimental error-induced hubs, a straightforward approach is to ignore all hub-associated links. This filtering method has been used to reduce the number of candidate pathways dramatically in the analysis of signal transduction networks [25]. However, the role of hub genes in the SGA data set is subtle. Genes in the *PAC10* complex are hubs that have enriched synthetic lethal interaction with genes in many other complexes, such as *CTF18* and *PAC11*. Many of the *PAC10*-associated links are biologically relevant since *PAC10* is indeed functionally coupled to a broad spectrum of biological pathways which themselves are functionally associated. Thus removing hub links entirely unsurprisingly leads to the loss of useful information and failure to detect some relevant pathways. GIMF treats this problem by permitting an increase in the parameter p_{bj} for hub library genes that are not part of the motif. Thus, we are able to extract biologically meaningful pathways by keeping the hub library genes whose impact is, however, automatically down-weighted.

This idea is tested on the Dynein-Dynactin gene pairs. Using GIMF we identified 24 Dynein-Dynactin pairs with the original SGA dataset. Then we tested GIMF on five filtered datasets generated by removing interactions with the top 5, 10, 15, 20 and 25 hub library genes, respectively. The corresponding fractions of interaction eliminated are 4.4%, 7.8%, 10.9%, 13.6% and 16%. With model parameters unchanged, GIMF recovers (18, 15, 10, 7, 1) Dynein-Dynactin pairs on the five datasets, respectively. The reduced coverage is expected from the removal of some biologically relevant hub links. However, a substantial number of those pairs are retained when interactions with the top 5 and 10 hub library genes are absent. These results suggest that a statistical method that explicitly models the skewed degree distribution is a better strategy for pattern discovery in the presence of hubs than using simple filtering techniques in conjunction with methods that do not take into account the hub effect.

The assumption in GIMF that the probability of an edge between a query and a library gene pair is proportional to the degree of the library gene works sufficiently well for the synthetic lethal interaction dataset. However, when extending the present model to other types of networks especially those with non-directional links, it would be beneficial to characterize the link probability in a subgraph based on local connectivities [26]. In this model, the link probability between a pair of genes depends on the degree of both genes. This allows us to consider each interaction in the context of its subgraph, thus has a good promise to extract motifs in power-law networks by their local deviations from randomness [27]. It would be interesting to integrate the local models into our algorithm in motif extraction of other interaction networks such as protein interaction networks.

Recently, Kelley et al. have integrated physical protein-protein interactions to dissect synthetic lethal gene pairs into between-pathway and within-pathway paradigms [14]. While the focus of our study is different from their work, GIMF has an interesting correspondence with their algorithm. The algorithm proposed by Kelley et al. to construct between-pathway or within-pathway model is essentially a local search procedure described by Sharan et al [28]. Starting from a seed node, nodes whose contributions to the current seed are maximal are added one at a time. The operation is repeated in a breadth-first search fashion so long as it increases the overall score of the subgraph. This is equivalent to maintaining a set of motif and non-motif nodes each with probability 1 and only the interaction between directly linked nodes are considered during the iteration. In contrast, GIMF maintains a probability of being in the motif set for each node, thus allowing all nodes to have contribution in each iteration during motif building. The assignment of a node to the motif ver-

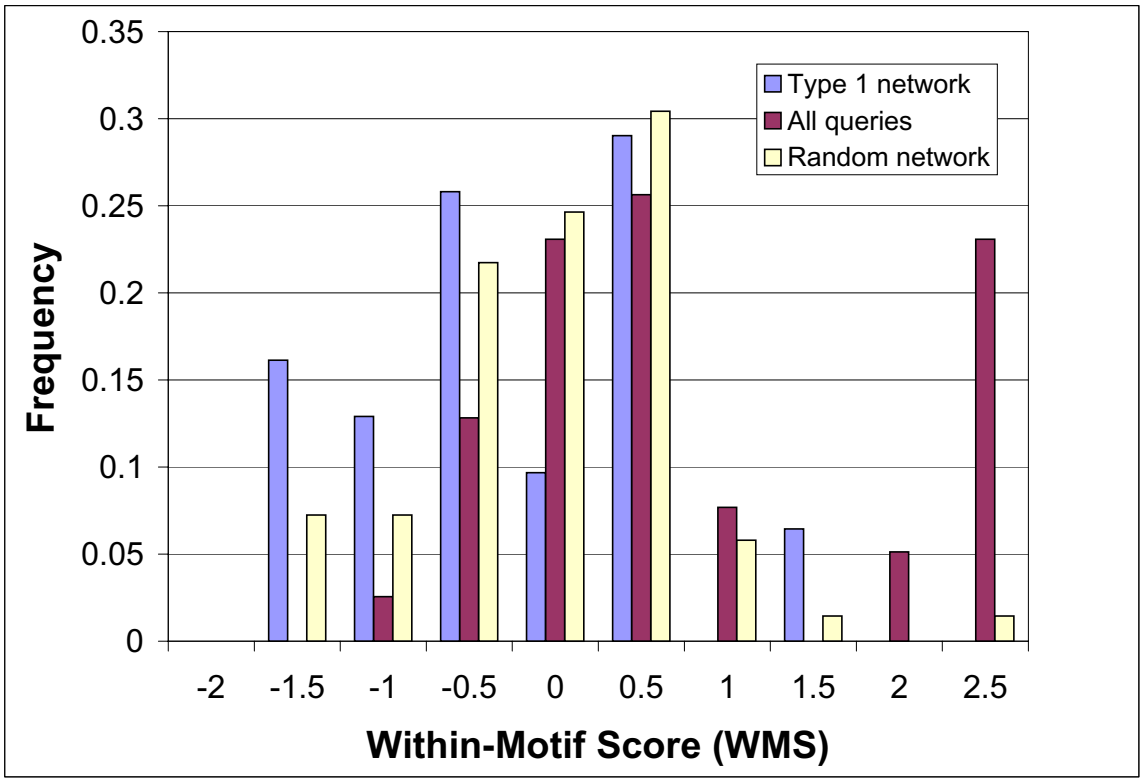


Figure 7
Within-Pathway Score (WMS) probability distributions. Distributions of the WMS on motifs from Type I network set (blue), all query set (red) and random network set (yellow). The WMS mean ± standard error for the Type I network is -0.250 ± 0.055; all query set, -0.011 ± 0.089 ; random network, 0.91 ± 0.21

sus non-motif category is only determined when the probabilities converge. A similar breadth-first search procedure can also be applied to GIMF in automatically extracting gene pathways. The between-pathway and within-pathway discovery by Kelley et al. [14] aligns with the conclusion by Tong et. al [12] that synthetic lethal interactions are more abundant between genes that have the same mutant phenotype and the genes encoding proteins within the same protein complex.

This idea is illustrated in Fig. 6, where "interaction", "no-interaction", "prohibited self-interaction" are represented by red, black and grey respectively. The matrix shows partial interaction profiles for five query genes. Query genes A, B, C and D, E belong to two different motifs. The within-pathway pattern shows the situation where synthetic lethal interactions are more abundant between motif members than between genes belonging to different motifs. The between-pathway pattern shows two motifs that represent two complementary pathways, with synthetic lethal interactions enriched between the pathways and depleted within a pathway. To permit a quantitative discussion, we define a within-motif score (WMS) to characterize whether synthetic lethal interactions for motif genes with each other are enriched (corresponding to the within-pathway pattern) or depleted (corresponding to the between-pathway pattern). Let WMS_i represent the score for motif *i* given by

$$WMS_i = \sum_{j=1}^{Q=126} Z_j^i \log(f_j^i / b_j^i) \Big/ \sum_{j=1}^{Q=126} Z_j^i \qquad (10)$$

where

$$f_j^i = \frac{(\text{Number of interactions with members of motif } i) + 1}{(\text{Total number of members of motif } i) - Z_j^i + 1} \text{ and}$$

$$b_j^i = \frac{(\text{Number of interaction partners}) + 1}{(\text{Total number of possible interaction partners}) + 1}. \qquad (11)$$

The total number of motif members is the denominator of Eq. 10, and the add-one pseudocounts in f_j^i and b_j^i bound the output of the log transform. The more negative an WMS, the more a motif reflects between-pathway interactions.

The WMS was computed for three sets of motifs generated: i) for seeds in the Type 1 network; ii) for all seeds from the query set; iii) for seeds in 100 randomized datasets described earlier (Fig. 7). The distribution of WMS values for motifs in the actual network appears bimodal, with greater probability for motifs with between-pathway character (WMS < 0). The WMS distribution for the Type 1 network has significantly more between-pathway character compared to motifs discovered in random network (one-sided, unequal variance t-test on WMS values, p-value = 1.4×10^{-5}). Motifs in the entire network also have significantly more between-pathway character, as judged by smaller WMS values, than motifs in the random network (p-value = 6.6×10^{-5}). Motifs from the Type 1 network show marginal significance for negative WMS values (one-sided z-test, p-value = 0.055), whereas motifs from the random network have significantly positive WMS values (p-value 4.5×10^{-6}). In summary, these results demonstrate that synthetic lethal interactions leading to motifs have significant between-pathway character, particularly when compared with motifs detected in randomized networks.

Though the purpose of this study is to develop a probabilistic model for characterizing synthetic lethal interaction motifs and a pathway identification algorithm based on synthetic lethal interaction datasets, the model holds good potential as an integrative method which combines multiple sources of evidence. If the sources of evidence are independent, the new likelihood function should be the multiplication of those for individual evidences. When the sources of evidence are not independent, then a Bayesian learning approach such as the framework developed by Jansen et al. [29] should be considered. A detailed discussion on the extension of GIMF into an integrative approach is however, beyond the scope of this study and hence will not be further considered here.

Conclusion

A probabilistic model and an automated algorithm (GIMF) have been shown to be effective in unsupervised motif learning of genetic interaction data. Starting from a seed pattern of genetic interaction partners, the method iteratively identifies genes that share the pattern and characterizes the pattern with a probabilistic motif. Functional associations are inferred from motif membership, rather than from existence of a direct genetic interaction linking two genes. Genes that belong to the same connected components in Type I and Type II networks have well correlated GO annotations, and are more likely to share annotations than genes connected by direct synthetic lethal interactions. Synthetic lethal interactions tend to be depleted between genes within a motif, suggesting that synthetic lethal interactions occur primarily between-pathway rather than within-pathway.

Several desirable features of the proposed algorithm for analyzing genetic interaction data include strong 0/1 predictions for genes sharing a motif, asymmetric property and the ability to automatically down-weight the impact of promiscuous genes with large degrees. We have shown that the asymmetry can be exploited to identify even tighter associations between genes and mask the impact of promiscuous genes. Furthermore, we conjecture that this asymmetric property may be useful in discriminating genes that are exclusive to a single pathway from genes that are shared in multiple pathways.

The probabilistic motifs naturally down-weight the importance of promiscuous genes with many interaction partners. When the roles of hubs are not purely due to experimental bias, it is more likely to retain biologically relevant information by modelling it probabilistically than by simple filtering. GIMF has an interesting correspondence with a log-odd score based approach. However, an important difference is GIMF performs a global search of a subgraph with best cohesiveness based on a seed. The computation of GIMF is highly efficient. It is well suited for building motifs around a subset of genes of interest with several choices of stringency.

Methods

Correlations for Gene Ontology (GO) annotation are computed for three categories: biological process, molecular function and cellular component (unpublished data, Ye et al). Within each category, the correlation coefficient is computed as follows:

Find the deepest level in GO hierarchy at which the pair of genes shares an annotation, which we denote by d.

Find the maximum and minimum value of d among all query gene pairs (i, j) where $i = 1,2,..., Q$ and $j = 1,2,...,Q$, Q is the total number of query genes.

The GO annotation correlation (biological process, molecular function and cellular components) for a pair of gene is defined by

$$correlation = \frac{d - d_{\min}}{d_{\max} - d_{\min}} \qquad (12)$$

Authors' contributions

YQ developed the GIMF model and carried out the data analysis. PY provided the code for the calculation of GO correlations and suggested the test of robustness against hub removal. JSB supervised the study.

Additional material

Additional File 1

Supplementary methods, figures, and tables.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-6-288-S1.pdf]

Additional File 2

Complete list of motif members for the 126 query genes using initialization p = 0.95.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-6-288-S2.txt]

Additional File 3

Tar file containing Matlab code, input, and output.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-6-288-S3.tar]

Acknowledgements

JSB acknowledges funding from the Whitaker Foundation and the NIH. YQ acknowledges support from Institute for Pure and Applied Mathematics for a relevant workshop, IBM for a Ph.D fellowship and Dr. Jianbo Gao for stimulating discussions.

References

1. Hartwell LH, Hopfield JJ, Leibler S, Murray AW: **From molecular to modular cell biology.** *Nature* 1999, **402(6761 Suppl):**C47-52.
2. Barabasi AL, Oltvai ZN: **Network biology: understanding the cell's functional organization.** *Nat Rev Genet* 2004, **5(2):**101-113.
3. Kitano H: **Computational systems biology.** *Nature* 2002, **420(6912):**206-210.
4. Xia Y, Yu H, Jansen R, Seringhaus M, Baxter S, Greenbaum D, Zhao H, Gerstein M: **Analyzing cellular biochemistry in terms of molecular networks.** *Annu Rev Biochem* 2004, **73:**1051-1087.
5. Gabaldon T, Huynen MA: **Prediction of protein function and pathways in the genome era.** *Cell Mol Life Sci* 2004, **61(7-8):**930-944.
6. Fraser AG, Marcotte EM: **A probabilistic view of gene function.** *Nat Genet* 2004, **36(6):**559-564.
7. Tong AHY, Evangelista M, Parsons AB, Xu H, Bader GD, Page N, Robinson M, Raghibizadeh S, Hogue CWV, Bussey H, Andrews B, Tyers M, Boone C: **Systematic genetic analysis with ordered arrays of yeast deletion mutants.** *Science* 2001, **294(5550):**2364-2368.
8. Ooi SL, Shoemaker DD, Boeke JD: **DNA helicase gene interaction network defined using synthetic lethality analyzed by microarray.** *Nat Genet* 2003, **35(3):**277-286.
9. Pan X, Yuan DS, Xiang D, Wang X, Sookhai-Mahadeo S, Bader JS, Hieter P, Spencer FA, Boeke JD: **A robust toolkit for functional profiling of the yeast genome.** *Submitted* 2004.
10. van Haaften G, Vastenhouw NL, Nollen EA, Plasterk RH, Tijsterman M: **Gene interactions in the DNA damage-response pathway identified by genome-wide RNA-interference analysis of synthetic lethality.** *Proc Natl Acad Sci U S A* 2004, **101(35):**12992-12996.
11. Baugh LR, Wen JC, Hill AA, Slonim DK, Brown EL, Hunter CP: **Synthetic lethal analysis of Caenorhabditis elegans posterior embryonic patterning genes identifies conserved genetic interactions.** *Genome Biol* 2005, **6(5):**R45.
12. Tong AHY, Lesage G, Bader GD, Ding HM, Xu H, Xin XF, Young J, Berriz GF, Brost RL, Chang M, Chen YQ, Cheng X, Chua G, Friesen H, Goldberg DS, Haynes J, Humphries C, He G, Hussein S, Ke LZ, Krogan N, Li ZJ, Levinson JN, Lu H, Menard P, Munyana C, Parsons AB, Ryan O, Tonikian R, Roberts T, Sdicu AM, Shapiro J, Sheikh B, Suter B, Wong SL, Zhang LV, Zhu HW, Burd CG, Munro S, Sander C, Rine J, Greenblatt J, Peter M, Bretscher A, Bell G, Roth FP, Brown GW, Andrews B, Bussey H, Boone C: **Global mapping of the yeast genetic interaction network.** *Science* 2004, **303(5659):**808-813.
13. Wong SL, Zhang LV, Tong AHY, Li ZJ, Goldberg DS, King OD, Lesage G, Vidal M, Andrews B, Bussey H, Boone C, Roth FP: **Combining biological networks to predict genetic interactions.** *P Natl Acad Sci USA P Natl Acad Sci USA* 2004, **101(44):**15682-15687.
14. Kelley R, Ideker T: **Systematic interpretation of genetic interactions using protein networks.** *Nature Biotechnology* 2005, **23(5):**561-566.
15. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, Remor M, Hofert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwein C, Heurtier MA, Copley RR, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G: **Functional organization of the yeast proteome by systematic analysis of protein complexes.** *Nature* 2002, **415(6868):**141-147.
16. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, Yang L, Wolting C, Donaldson I, Schandorff S, Shewnarane J, Vo M, Taggart J, Goudreault M, Muskat B, Alfarano C, Dewar D, Lin Z, Michalickova K, Willems AR, Sassi H, Nielsen PA, Rasmussen KJ, Andersen JR, Johansen LE, Hansen LH, Jespersen H, Podtelejnikov A, Nielsen E, Crawford J, Poulsen V, Sorensen BD, Matthiesen J, Hendrickson RC, Gleeson F, Pawson T, Moran MF, Durocher D, Mann M, Hogue CW, Figeys D, Tyers M: **Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry.** *Nature* 2002, **415(6868):**180-183.
17. Lawrence CE, Reilly AA: **An Expectation Maximization (EM) Algorithm for the Identification and Characterization of Common Sites in Unaligned Biopolymer Sequences.** *PROTEINS: Structure, Function, and Genetics* 1990, **7:**41-51.
18. Bailey TL: **Unsupervised Learning of Multiple Motifs in Biopolymers Using Expectation Maximization.** *Machine Learning Journal* 1995, **21:**51-83.
19. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC: **Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment.** *Science* 1993, **262(5131):**208-214.
20. Dempster AP, Laid NM, Rubin DB: **Maximum likelihood from incomplete data via the EM algorithm.** *J Royal Statistical Soc B* 1977, **39:**1-38.
21. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U: **Network motifs: simple building blocks of complex networks.** *Science* 2002, **298(5594):**824-827.
22. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y: **A comprehensive two-hybrid analysis to explore the yeast protein interactome.** *Proc Natl Acad Sci U S A* 2001, **98(8):**4569-4574.

23. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM: **A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae.** *Nature* 2000, **403(6770):**623-627.
24. Ye P, Peyser B, Pan X, Boeke JD, Spencer FA, Bader JS: **Quantified measures of systems robustness in yeast.** Baltimore ; 2004.
25. Steffen M, Petti A, Aach J, D'Haeseleer P, Church G: **Automated modelling of signal transduction networks.** *BMC Bioinformatics* 2002, **3(1):**34.
26. Itzkovitz S, Milo R, Kashtan N, Ziv G, Alon U: **Subgraphs in random networks.** *Phys Rev E Stat Nonlin Soft Matter Phys* 2003, **68(2 Pt 2):**26127.
27. Berg J, Lassig M: **Local graph alignment and motif search in biological networks.** *Proc Natl Acad Sci U S A* 2004, **101(41):**14689-14694.
28. Sharan R, Suthram S, Kelley RM, Kuhn T, McCuine S, Uetz P, Sittler T, Karp RM, Ideker T: **Conserved patterns of protein interaction in multiple species.** *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102(6):**1974-1979.
29. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M: **A Bayesian networks approach for predicting protein-protein interactions from genomic data.** *Science* 2003, **302(5644):**449-453.